Current Topics in
Virology

Review

Recent advances in human cytomegalovirus genomics

Steven Sijmons, Marc Van Ranst and Piet Maes*
Laboratory of Clinical Virology, Rega Institute for Medical Research, KU Leuven, Belgium

ABSTRACT

It has been 22 years since the first human cytomegalovirus (HCMV) full genome sequence was published. Together with the first sequence also came the first prediction of the coding capacity of this intriguing herpesvirus, harbouring the largest genome of all human-infecting viruses. Through the years, the HCMV genome map was refined, but it is still far from complete. Comparisons of the genomes of highly passaged and attenuated laboratory strains with those of recent clinical isolates have shown the former underwent substantial rearrangements with gene deletions and duplications. It became clear that different clinical isolates display remarkable genetic variability in large portions of their genomes. This led to a quest for clinical correlates of different genetic variants. Until now, this approach did not give satisfactory results and it is still not clear if certain genetic variants are associated with distinct disease outcome. More recently, extra levels of complexity have been added to the existing picture of HCMV genetic variability. Infections with multiple HCMV strains seem to be very common and the dynamics of different strains show unpredictable and stochastic behaviour. Deep sequencing efforts have unveiled an extensive intrapatient genetic variability, which was unexpected for a dsDNA virus. Recent efforts to elucidate the HCMV transcriptome have painted a sophisticated picture with lots of splicing and alternative transcripts. It is clear that we are only unveiling the tip of the iceberg regarding HCMV genomics and transcriptomics and their role in the myriad viral functions. Recent advances in technology and bioinformatics provide us with the tools to begin to tackle these interesting questions.

KEYWORDS: cytomegalovirus, human herpesvirus 5, genomics, transcriptomics, full genome

ABBREVIATIONS

HCMV	:	Human cytomegalovirus
dsDNA	:	double-stranded DNA
UL	:	unique long
US	:	unique short
TRL/IRL	:	terminal/internal repeat long
IRS/TRS	:	internal/terminal repeat short
ORF	:	open reading frame
CCMV	:	chimpanzee cytomegalovirus
UL128L	:	UL128 locus
IGA	:	Illumina Genome Analyzer
NGS	:	next-generation sequencing
miRNA	:	microRNA

INTRODUCTION

HCMV is the prototype member of the herpesvirus subfamily *Betaherpesvirinae* [1]. These viruses are characterized by lengthy replication cycles, are strictly cell-associated *in vitro* and do generally not cross host species barriers. HCMV is distributed worldwide and infections are common, with seroprevalences of 45-100% in women of reproductive age, depending on geographic location, socio-economic background and age [2]. Infections of healthy children and adults are usually

*Corresponding author: Piet Maes,
Laboratory of Clinical Virology,
Minderbroedersstraat 10,
BE-3000 Leuven, Belgium.
pmaes3@uzleuven.be

asymptomatic but the virus establishes a lifelong persistence as a latent infection, from which it can reactivate to spread infectious progeny. HCMV can cause serious sequelae during infection of immunocompromised individuals and fetuses [3]. Worldwide, HCMV is the most common congenital infection, with an overall birth prevalence of 0.64% [4].

Like other herpesviruses, HCMV has a characteristic virion morphology consisting of a linear dsDNA genome wrapped in an icosahedral capsid, an amorphous tegument layer and a lipid bilayer envelope. The HCMV genome has a size of approximately 235 kbp, which makes it the longest genome of all human viruses. It contains two unique regions, termed unique long (UL) and unique short (US), both flanked by a pair of inverted repeats ab (TRL/IRL) and ac (IRS/TRS). The overall genome arrangement can be described as ab-UL-b'a'c'-US-ca where primes designate inverted orientations (Fig. 1). UL and US can be inverted relative to each other giving rise to four genomic isomers [5].

In this review, we want to give an overview of our understanding of the HCMV genome, its coding capacity and diversity, which has developed over the past 20 years. The recent progress made through the application of novel high-throughput sequencing technologies will be highlighted.

HCMV genome sequencing and prediction of coding capacity

The first complete genome sequence of HCMV was published in 1990 by Chee and colleagues [6]. The sequence was derived through M13 shotgun cloning and chain-termination sequencing of the extensively passaged laboratory strain AD169, which was and is highly used in HCMV research. At the time, it was the largest contiguous sequence ever derived. The authors calculated that the effort to sequence this single genome corresponded to one person working for 12 years on the project [7]. A first genetic map of HCMV was derived, containing 208 open reading frames (ORFs) that were thought to encode proteins. 14 of these ORFs were duplicated in the TRL/IRL repeats. Like other herpesviruses, most of the HCMV genes are positioned end-to-end, with little non-coding sequence. A striking feature of the HCMV genome map was the apparent duplication and divergence of a large set of genes, which were grouped in gene families. 52 of the predicted ORFs could be grouped into 9 gene families.

Comparisons of genome regions of AD169 with Towne and Toledo strains revealed major genome rearrangements [8, 9]. The highly passaged and low virulent AD169 and Towne strains, which had been developed as vaccine candidates, were missing a segment of 15 and 13 kbp respectively which was present in the Toledo strain (Fig. 1). Toledo was passaged significantly less (exact passage numbers are not known) and had produced clinically apparent disease in clinical trials. These missing segments were located at the junction of the UL and b' (IRL) regions. Total genome length did not change significantly, since the b repeats were much smaller in Toledo. Inspection of five additional clinical isolates showed that they all contained the 15 kbp region missing in AD169. Apparently, the extensive passaging and loss of virulence of AD169 and Towne had been

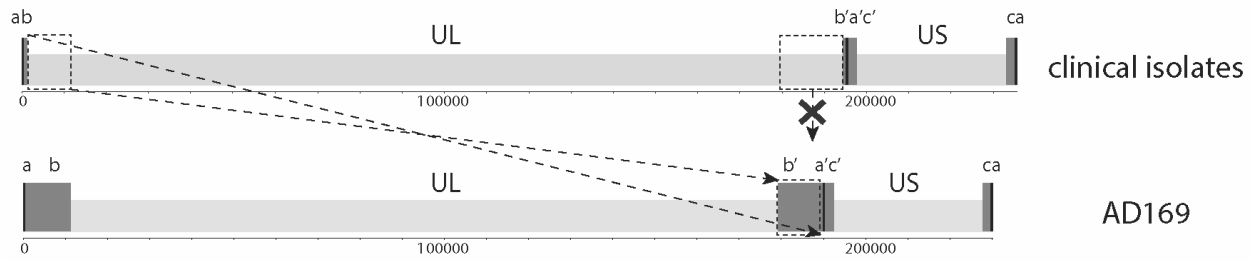

Fig. 1. Schematic representation of the overall genome rearrangement of HCMV clinical isolates and highly passaged laboratory strains, exemplified by strain AD169. The deletion and compensatory duplication that arose during extensive passage are indicated. ab corresponds to TRL, b'a' to IRL, a'c' to IRS and ca to IRL.

accompanied by the loss of a region in the 3' end of UL. An inverted copy of a region in the 5' end of UL had replaced this region effectively enlarging the b repeats. This compensatory duplication is probably a consequence of the genome length restrictions that are imposed by the packaging process and virion architecture. The 15 kbp region (commonly referred to as the UL/b' region) was predicted to encode 19 additional ORFs. While these ORFs are evidently dispensable for growth in fibroblasts, they immediately attracted interest as possible modulators of cell tropism, immune regulation and latency.

The initial prediction of 208 coding ORFs by Chee *et al.* [6] was effectively based on the criteria of a minimum polypeptide length of 100 amino acids and a maximal overlap of 60%. The authors admitted these criteria could exclude smaller or highly spliced functional ORFs. Furthermore, it was unsure whether all predicted ORFs encoded *bona fide* proteins. Different complementary approaches were taken to achieve more accurate predictions of the HCMV coding capacity. In a first approach, the genome sequence of chimpanzee cytomegalovirus (CCMV), the closest known relative of HCMV, was determined and compared with the AD169 sequence complemented by the Toledo UL/b' region [10]. The rationale of this method is that functional ORFs are conserved during evolution so that ORFs conserved between HCMV and CCMV are very likely functional. The authors concluded 51 of the proposed ORFs are unlikely to encode proteins, modified 24 other ORFs and defined 10 novel ORFs. A preliminary estimate of 164 to 167 genes for wild-type HCMV was proposed. In a similar approach, the coding potential of AD169 was reassessed using a gene-finder algorithm that predicts the likelihood of the functionality of ORFs through matching with a database of amino acid patterns [11]. Orthologues of these predicted ORFs were then searched in the genomes of chimpanzee, rhesus and murine cytomegalovirus. This study proposed an estimate of 192 unique ORFs that are potentially encoding and results were generally in agreement with the study based on the CCMV comparison. The same authors also took a different approach by sequencing two laboratory strains and four briefly passaged clinical isolates after cloning them as bacterial artificial chromosomes [12]. They reported a set of 252 ORFs that were conserved in all four clinical isolates including a set of 29 that were previously unrecognized. Comparison of the conserved ORFs of the four clinical isolates also revealed that while many of the ORFs were highly conserved, a subset displayed remarkable variability among different strains. These variable ORFs were found scattered throughout the HCMV genome. Dolan *et al.* determined the complete genome sequence of a minimally passaged isolate (passage 3) from a congenitally infected infant, termed strain Merlin [13]. Additionally, they studied substantial regions of the genome in other isolates, including unpassaged strains. They also found high variation in a subset of genes, many of which encode secreted and membrane-associated proteins. Based on these combined data, the authors refined previous annotations and defined a set of 165 genes reflecting the wild-type HCMV coding capacity. The sequence of strain Merlin has become the reference sequence for wild-type HCMV and its annotation is constantly updated incorporating new insights. The genetic map of HCMV as currently annotated on the HCMV reference sequence (NC_006273) contains 170 genes and is presented in Fig. 2.

While large genome rearrangements have occurred in highly passaged laboratory strains like AD169 and Towne, these strains also show several mutations affecting other genes than the ones in the UL/b' region. AD169 is mutated in RL5A, RL13, UL36 and UL131A, whereas Towne displays frameshifts in RL13, UL1, UL130, US1 and US9 [14]. In fact, the Toledo strain that led to the discovery of the large deletion in AD169 and Towne itself had an inversion in the UL/b' region, resulting in disruption of UL128 [8, 9]. Furthermore, Toledo is mutated in RL13 and UL9. Even the minimally passaged Merlin strain has a mutation in UL128 [13]. Most strains passaged in fibroblast cell lines seem to contain mutations in 1 of 3 neighbouring genes, UL128, UL130 and UL131A, the so-called UL128 locus (UL128L). The products of this gene locus form a complex with the viral glycoproteins gH and gL that is required for entry into endothelial and epithelial cells, but not for entry into fibroblasts [15]. When developing a

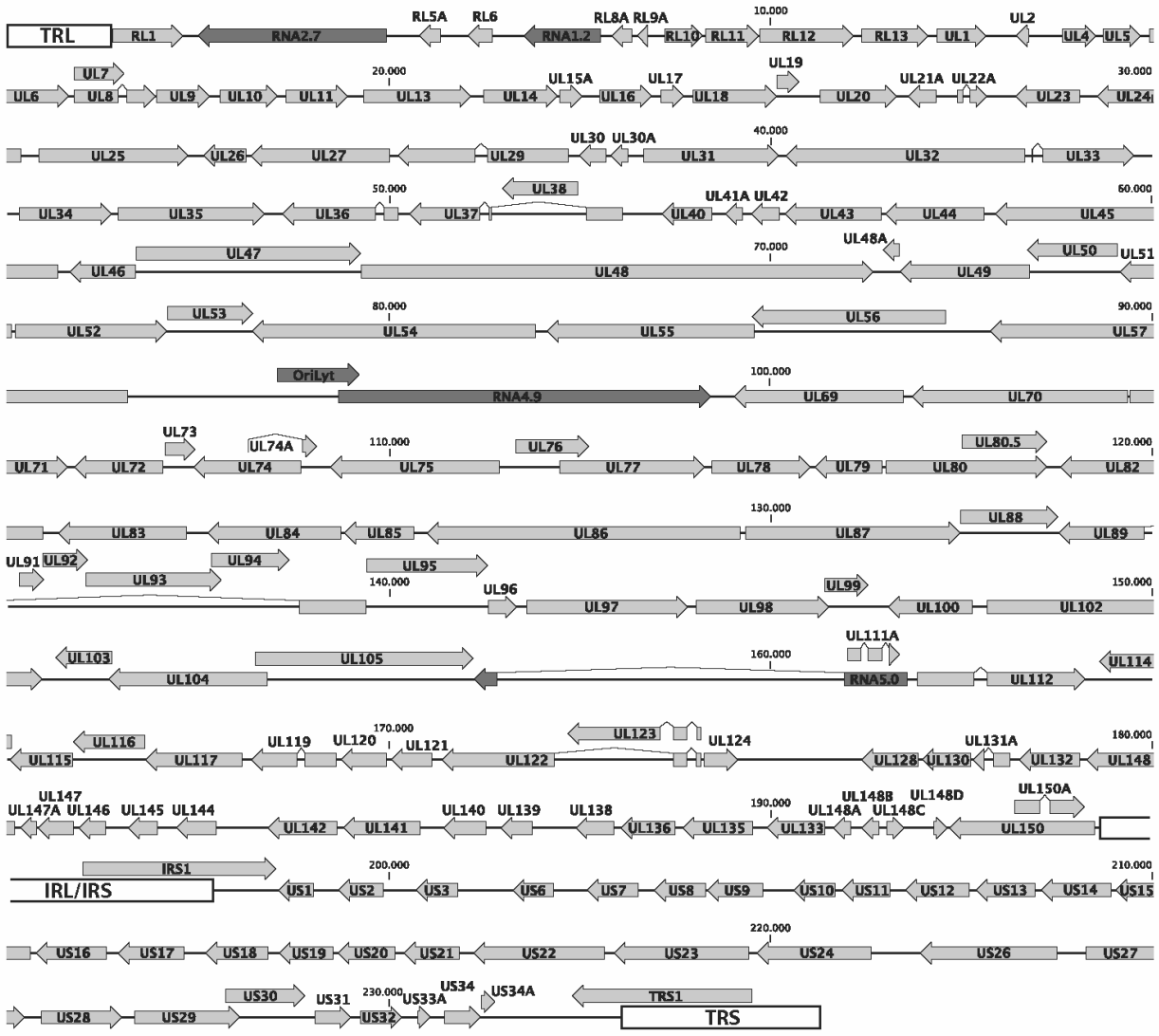

Fig. 2. Genetic map of wild-type HCMV. This map is based on the current annotation of reference strain Merlin (NC_006273). Genes are designated in light gray, large non-coding RNAs and the lytic origin of replication (oriLyt) in dark gray.

BAC vector of the Merlin strain, it was noted that it also contained a defective RL13 gene [16]. The authors went back to the original Merlin isolate and identified that, while the RL13 consensus sequence was wild-type, different clones all contained mutations at different positions in RL13. When RL13 was repaired, it repressed HCMV replication in various cell types and RL13 mutants rapidly and invariably re-emerged. Repair of UL128L only repressed replication in fibroblasts and mutations were again acquired, although somewhat less rapidly than for RL13. Dargan *et al.* elaborated on this by characterizing mutational dynamics during passaging of four clinical strains in fibroblasts, epithelial and endothelial cells [17]. They confirmed the invariable mutation of RL13 in all cell types and UL128L in fibroblasts and demonstrated the suppressive effects on growth in fibroblasts were at least partially independent. Additionally, they identified several other regions of the genome that mutated in some but not all strains. These results showed that all clinical isolates were genetically unstable during passage in different cell types. Strain sequences determined were still the product of PCR-based amplification and traditional Sanger sequencing.

In fact, this study somehow marks the end of an era in HCMV genomic research, because simultaneously the first reports were being published making use of a completely new generation of sequencing technologies.

Next-generation sequencing in HCMV genomics

The advent of new, massively parallel sequencing technologies in the second half of the past decennium opened up new possibilities in high-throughput and deep sequencing [18]. These technologies, of which the 454 (Roche) and Illumina (Solexa) platforms were the front-runners, generate an enormous amount of short sequence reads from minimal input materials without the need for laborious cloning procedures. The newly emerging high-throughput sequencing technologies were first applied to HCMV genomics in 2009. The Illumina Genome Analyzer (IGA) (Solexa) was used to elucidate the presence of two variants in preparations of the Towne strain and identified a new, more genetically intact variant of AD169 called varUC [14].

In a next step, the usefulness of next-generation sequencing (NGS) for sequencing the complete genomes of clinical HCMV isolates was evaluated using IGA, compared to a PCR-based amplification and Sanger sequencing approach [19]. Isolates were sequenced both after cell culture passage and directly from clinical material. While the PCR-based approach was successful, even with clinical samples containing only 3% of HCMV DNA, it proved to be very laborious. IGA sequencing provided a much more efficient alternative for sequencing HCMV strains directly from clinical isolates, although assembly of the enormous amount of NGS short sequence reads was not straightforward. Due to the inherent variability of HCMV isolates, direct mapping of NGS reads on a reference sequence did not yield satisfactory results. In stead, the authors devised an approach in which NGS reads were initially assembled without the use of a reference sequence, i.e. *de novo* assembly. The sequences produced via this method were then used to assist reference-dependent assembly of the original NGS reads. This approach proved to be successful in determining the sequence of clinical HCMV isolates, even from unamplified material containing only 3% HCMV DNA. It is, however, clear that some enrichment or amplification step is necessary to make the sequencing of HCMV strains directly from clinical material achievable in high-throughput. This study confirmed the fact that all passaged strains contained one or more mutations. Furthermore, some evidence did point at a possible presence of mutations in the original clinical isolates. Some mutations were indeed shared between samples from independent clinical settings and one of the unpassaged strains was mutated in RL5A and UL111A.

Another NGS platform, the 454 GS-FLX (Roche) was used to determine the first full genome sequence of an Asian HCMV strain [20]. This strain was plaque purified before a short passage on fibroblasts and contained two mutated genes (UL1 and UL119).

While these previous studies focused on the determination of a consensus sequence from a viral population, NGS also offers new possibilities in deep sequencing of viral populations. NGS was first utilized to characterize HCMV populations by Görzer *et al.* [21]. Subsequently, Renzette *et al.* took this approach to a genome-wide perspective [22]. This added three HCMV full genome sequences, or major genome types as the authors refer to them, to the collection on GenBank. Details of these two studies are elaborated in the corresponding part about intrahost diversity. All HCMV full genome sequences currently published are presented in Table 1.

Interhost diversity and clinical significance

Like stated earlier, full genome comparisons of HCMV strains led to the understanding that substantial regions of the genome were highly variable between independent strains [12, 13]. It was already known that some of the genes encoding viral glycoproteins are not genetically homogeneous and that these polymorphisms exist as highly diverged clusters of alleles, so called genotypes [23-25]. Some of the genes in the UL/b' region also attracted interest in this regard [26-28]. In Fig. 3, the evolutionary divergence of all genes on the HCMV genome is presented, based on a set of 39 full genome sequences of clinical HCMV strains. Next to some of the genes

Table 1. HCMV full genome sequences available on GenBank.*

GenBank Id	Strain name	Source	Passage history	Sequence length (bp)	Reference
NC_006273° GU179001	Merlin	urine from a congenitally infected infant	passaged 3 times in human fibroblasts	235646	[13]
BK00039`	AD169 varUK	adenoids of a 7-year old girl	passaged extensively in human fibroblasts	230290	[6, 74]
FJ527563	AD169 varUC	adenoids of a 7-year old girl	passaged extensively in human fibroblasts	231781	[14]
AC146999	AD169-BAC	adenoids of a 7-year old girl	BAC clone from a plaque-purified AD169 derivative	233739	[12]
AY315197	Towne	urine of a 2-month-old infant with microcephaly and hepatosplenomegaly	passaged extensively in human fibroblasts	222047	[75]
AC146851	Towne-BAC	urine of a 2-month-old infant with microcephaly and hepatosplenomegaly	BAC clone from a plaque-purified Towne derivative	229483	[12]
FJ616285	Towne	urine of a 2-month-old infant with microcephaly and hepatosplenomegaly	passaged extensively in human fibroblasts	235147	[14]
GU937742	Toledo	urine from a congenitally infected infant	passaged several times in human fibroblasts	235398	-
AC146905	Toledo-BAC	urine from a congenitally infected infant	BAC clone from a plaque-purified Toledo derivative	226889	[12]
AC146907	FIX-BAC	cervical secretions of a pregnant woman with a primary HCMV infection	BAC clone from the VR1814 isolate	229209	[12]
AC146904	PH-BAC	transplant patient with HCMV disease	BAC clone from the PH isolate (passaged less than 12 times)	229700	[12]
AC146906	TR-BAC	AIDS patient with CMV retinitis	BAC clone from the TR isolate	234881	[12]
EF999921	TB40/E clone TB40-BAC4	throat wash of a bone marrow transplant recipient	BAC clone from TB40/E passaged 5 times in human fibroblasts and 22 times in human endothelial cells	229050	[76]

Table 1 continued..

GQ221974	3157	urine from a congenitally infected infant	passaged 3 times in human fibroblasts	235154	[19]
GQ466044	3301	urine from a congenitally infected infant	unpassaged	235703	[19]
GU179291	AF1	amniotic fluid	unpassaged	235937	[19]
GQ221973	HAN13	bronchoalveolar lavage	passaged 3 times in human fibroblasts	236219	[19]
GQ396663	HAN20	bronchoalveolar lavage	passaged 2 times in human fibroblasts	235728	[19]
GQ396662	HAN38	bronchoalveolar lavage	passaged 2 times in human fibroblasts	236112	[19]
GQ221975	JP	post mortem prostate tissue from an AIDS patient	unpassaged	236375	[19]
GU179288	U8	urine from a congenitally infected infant	unpassaged	235709	[19]
GU179290	U11	urine from a congenitally infected infant	unpassaged	234732	[19]
GU179289	VR1814	cervical secretions of a pregnant woman with a primary HCMV infection	unpassaged	235233	[19]
HQ380895	JHC	blood from a bone marrow transplant patient	plaque-purified and passaged 3 times in human fibroblasts	235476	[20]
JN379814”	U01	urine from a congenitally infected infant	unpassaged	232216	[22]
JN379815”	U04	urine from a congenitally infected infant	unpassaged	233910	[22]
JN379816”	U33	urine from a congenitally infected infant	unpassaged	232889	[22]

* transgenic strains are not included
° RefSeq strain
` adds 929 basepairs (region of UL42 and UL43) that are missing from the original sequence (X17403)
’’ major genome types containing gaps

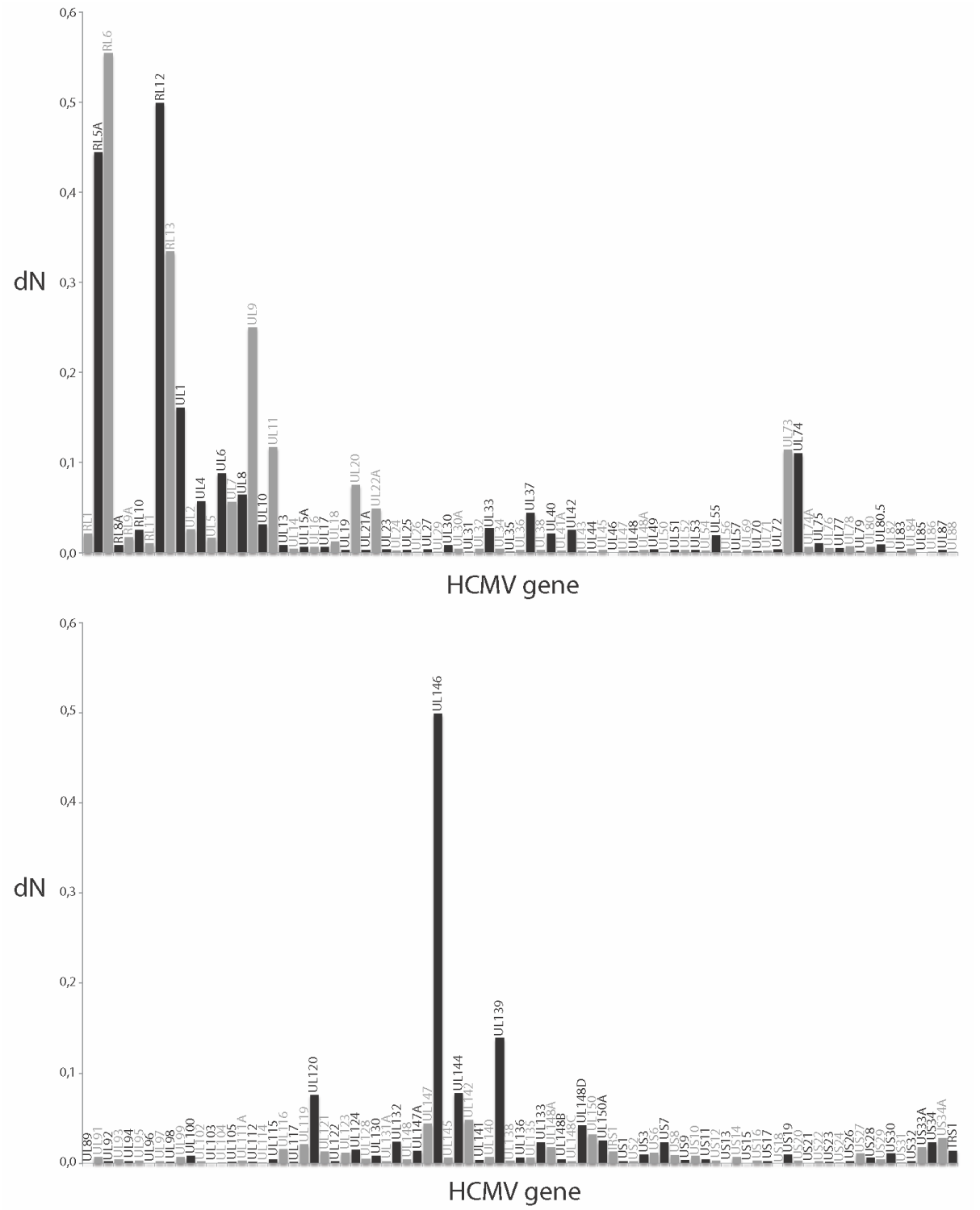

Fig. 3. Average codon-based evolutionary divergence of HCMV genes. The average number of non-synonymous substitutions per non-synonymous site dN was calculated for each HCMV gene based on 39 full genome sequences of recent clinical isolates. These included strains Merlin, 3157, 3301, AF1, HAN13, HAN20, HAN38, JP, U8, U11, VR1814, JHC, U01, U04 and U33 (Table 1) and 24 unpublished, in-house derived full genome sequences. Analyses were conducted using the Nei-Gojobori model [72]. Divergences were calculated using MEGA5 [73].

encoding viral glycoproteins (mainly gN/pUL73 and gO/pUL74) and genes in the UL/b' region (most notably UL146 and UL139), the high variability of the genes at the 5' end of the UL region is remarkable. These genes form the RL11 family, characterized by the presence of the RL11 domain [29]. Several studies already reported about the clustered nature of polymorphisms in this region of the HCMV genome, but although roles in cell tropism and immune evasion have been proposed for some members, functions of the RL11 family are still poorly characterized [30-32].

Considering the high variability in some of the HCMV genes, it is quite remarkable that HCMV genotypes seem to be stable, both within a person as within the population over time [27, 33]. The same genotypic sequences have been found in distinct geographic regions and until now there is no proof of genotypes associated with specific locations [34-36]. Apparently, the selective forces that have shaped this variability are not active anymore but exerted their effects a long time ago. The current hypothesis states that genotypes were formed through immune selection in populations of early humans or their predecessors, have been shaped by founder and bottleneck events, and have been distributed by the worldwide migrations of human populations in more recent times [35, 37, 38]. Moreover, genetic linkages between different HCMV loci seem to be very rare, mainly limited to regions in close proximity to each other [27, 32, 39]. Recombination has probably played an important factor in HCMV evolution and has generated a virtually infinite number of strains by combining genotypes at different sites [39-41].

A plethora of studies have been carried out to investigate potential correlations between viral genotypes and pathogenicity, focusing on genes encoding viral glycoproteins and immune modulatory proteins [42-45]. These studies were mainly motivated by the prospect of finding viral prognostic markers for severity of infection, which could help tailor medical interventions. We will not elaborate further on them here and refer to some excellent reviews that cover this matter extensively [37, 46, 47]. From the collected data, we can conclude that potential relationships between HCMV genotypes and pathogenicity remain elusive, which is not surprising, considering that studies usually have focused on one or two gene products. Taking into account the number of HCMV genes and the substantial variability in a large subset of them, establishing any viral prognostic marker will need a more comprehensive approach taking into account a larger set of variable genes, if not full genome sequences. With the current progress in high-throughput sequencing technologies, this kind of endeavour could become feasible in the near future.

Multiple infections and intrahost diversity

The search for viral determinants of pathogenicity gets even more complicated when we have a look at HCMV diversity within an individual host. In recent years, it has become increasingly clear that infections with multiple HCMV strains are common, both in immunocompromised and immunocompetent individuals (reviewed in [37]). The question whether these multiple infections are mostly the product of transmission of multiple strains at the same time or a consequence of sequential infection events, remains unresolved, but it seems that both are possible [48, 49]. It has been proposed that mixed infections could be disadvantageous for the patient, possibly through *trans*-complementation between strains [50]. This disadvantage is supported by studies looking at the effect of strain multiplicity on outcome in transplant recipients [51-54].

NGS technologies are particularly well suited for studying the dynamics of mixed viral populations, since the vast amount of NGS reads ensure multiple coverage of the same region of the genome or gene of interest. Görzer *et al.* made use of the 454 GS-FLX platform (Roche) to sequence amplicons of hypervariable HCMV genes UL73 (gN), UL74 (gO) and UL139 in lung transplant recipients [21]. They observed complex mixtures of up to six genotypes in one patient, with maximally 2 types making up the majority of the population. Comparison of serial samples of the same patient showed substantial changes in the abundance of genotypes over time. It is far from clear what processes shape this differential replication of individual strains in multiple infections.

While genotypes at individual loci can give an idea of intrahost strain diversity, much more

information can be extracted from a genome-wide perspective. Renzette *et al.* used the throughput of the Illumina GAII platform (Solexa) to sequence amplicons covering the complete genomes of three congenitally infected infants [22]. To distinguish true HCMV variants from PCR- and/or sequencing errors, they developed an error-filtering algorithm based on the resequencing of BAC clones of AD169 and Toledo. Mismatches in the sequences of the BAC clones were considered errors and used to establish filtering thresholds for the variants in the unknown samples. Using these filtering protocols, the authors found the HCMV populations in infants to have genetic diversity estimates comparable to those of RNA viruses like HIV and dengue virus, which was a surprising result for a dsDNA virus encoding a polymerase with proofreading capacity. These populations consisted of one highly abundant type and a high number of low-level variants, which is essentially in agreement with the findings of Görzer *et al.* [21]. Because of the fractional nature of NGS reads, it is impossible to infer whether highly abundant types at different locations actually form a single genome-wide type. At the moment, it is not clear how this diversity is generated and what role these low-level variants play in natural infections. They could provide the virus with a means of rapidly evolving to changes in its environment, but could also prove to be defective variants that are just a by-product of productive viral replication. More research is warranted to confirm these estimates of viral population diversity in other clinical and non-clinical settings. In this regard, the authors already applied their diversity estimation algorithm to HCMV populations in malignant gliomas and found these to be significantly less diverse [55].

It would also be interesting to elucidate if and how selection is acting on these populations. In this regard, Renzette *et al.* already provided some preliminary results that could point to positive selection acting on the intrahost populations [22]. This is obviously in contrast to the stability and negative selection that seem to shape genotypes in interhost populations [35]. The authors provide the hypothesis of a balance between both where intrahost diversity is promoted as an evolutive advantage in an ever-changing environment and negative selection prevents these mutants from reducing the overall fitness of the population [22]. Further elucidation of the forces that are shaping HCMV viral populations in different patient groups is necessary to improve our understanding of the evolutionary dynamics of this large dsDNA virus.

HCMV transcriptomics

For many protein-coding ORFs, prediction by comparative genomic and pattern-based similarity approaches as described earlier has been confirmed by information on the expression and function of their products. A large fraction of predicted genes remain, however, speculative and their transcriptional profile has not been studied in detail. While ORF-searching algorithms based on genomic DNA sequences certainly have their virtue in deciphering the coding capacity of an organism, there is mounting evidence that these approaches yield genomic maps that are an oversimplification of reality. An extra level of complexity only becomes apparent when transcriptional patterns are studied. For HCMV, some complex transcriptional processes have been discovered (reviewed in [56]): multiple transcripts sharing common 5’ or 3’ ends, complex and adaptable splicing patterns, antisense transcription and transcription of non-coding and miRNAs.

When cloning and sequencing cDNA libraries derived from HCMV transcripts, Zhang *et al.* observed that a striking 45% of clones came from genomic regions thought to be non-coding [57]. Perhaps even more surprising, 55% of clones were derived from transcripts antisense to 56 known or predicted ORFs. Through the use of NGS tools, Gatherer *et al.* brought full genome resolution to HCMV transcriptome studies [58]. They confirmed the presence of antisense transcription throughout the genome, although generally transcribed at a lower level than their sense counterpart and overall accounting for 8.7% of transcription. The observed abundance of antisense transcription leads to the hypothesis that this could be a previously unrecognized viral mechanism for gene regulation. This idea is fostered by the growing understanding of the roles of antisense transcription as an important epigenetic factor in other pro- and eukaryotes (reviewed in [59]). At the moment, the role of

antisense transcripts in HCMV infections is, however, still largely uncharacterized.

Next to the potential role of antisense transcripts, the full genome HCMV transcriptome study provided some other important insights [58]. It was observed that 65.1% of viral transcripts were involved in the production of four non-coding RNAs (RNA2.7, RNA1.2, RNA4.9 and RNA5.0) (Fig. 2). These RNAs do not overlap substantially with known ORFs and are therefore thought not to function through translation. A staggering 46.8% of viral transcripts were dedicated to the production of RNA2.7 that has been shown to regulate mitochondria-induced cell death thereby inhibiting apoptosis [60]. This large involvement of antisense and non-coding RNAs reciprocally implied that only one-third of viral polyadenylated RNAs were translated into proteins. Furthermore, a large number of unrecognized splicing sites were discovered and confirmed through additional experimentation, unveiling splicing is much more common and complex than previously recognized. Splice sites are now characterized in about one-third of ORFs and while necessary for expression in some, they could have more subtle regulatory roles in other. Altogether, this study revealed HCMV transcription to be much more complex than expected, providing the virus with a wealth of tools to regulate its functions. The transcriptome profiling also led to the designation of four new genes (RL8A, RL9A, UL150A and US33A), bringing the estimate of HCMV genes to 170 (Fig. 2). This indicates the added value of transcriptomics in genome annotation.

The complexity of the HCMV coding capacity was recently confirmed by characterization of the translation products that arise during infection [61]. The authors used a technique called ribosome profiling to generate libraries of ribosome-protected mRNA fragments that were sequenced using Illumina NGS platforms. Examination of ribosome footprints led to the identification of 751 translated ORFs with only 147 previously predicted to be coding. These included ORFs positioned within existing ORFs (both in-frame and out of frame), short ORFs upstream of existing ORFs, antisense ORFs and short ORFs lying within regions that were previously not predicted to be coding. Furthermore, multiple ORFs were translated from the long non-coding RNAs 1.2, 2.7 and 4.9. A large fraction of these ORFs were confirmed through high-resolution tandem mass spectrometry or tagging approaches. The authors also stress the role of alternative 5' ends in the temporal regulation of viral gene expression. By changing the 5' end of transcripts during the course of infection, the virus can express different proteins from overlapping coding regions in a temporal fashion.

Next to the longer non-coding RNAs, it was shown that viruses also encode microRNAs (miRNAs) [62]. MicroRNAs are small RNAs (approx. 23 nt) that play a part in post-transcriptional repression of gene expression through targeting of protein-coding mRNAs [63]. MiRNAs enable regulation of viral and cellular genes while being non-immunogenic and only requiring minimal space in the viral genome. The HCMV genome is known to contain miRNA precursors that are processed to mature miRNAs during infection [64, 65] (reviewed in [66]). Stark *et al*. recently studied the profile of small RNAs (both viral and host-derived) during HCMV infection through NGS analysis (IGA) [67]. They found that 20% of the small RNA population encoded viral miRNAs and designated two novel HCMV miRNAs. At the moment, 22 mature HCMV miRNAs are characterized. The authors provided evidence that all HCMV miRNAs are incorporated into the endogenous host silencing machinery demonstrating their functionality. Furthermore, novel small viral RNAs were identified that were distinct from miRNAs. These were especially observed across long non-coding RNAs like RNA2.7. While these could be degradation products from this highly transcribed region of the HCMV genome, other highly transcribed regions were not found to produce these small RNAs. These results are in agreement with the observations made by Stern-Ginossar *et al.* in the ribosome profiling study [61]. Regarding their ubiquity, these small RNAs (and their putative protein products) could play important regulatory roles during infection but up to now there is no data available to strengthen this hypothesis.

There are some caveats to the wealth of data that are emanating from these HCMV transcriptome studies. As long as these RNA characterizations

are not complemented by data surrounding the functionality of the predicted RNAs, their importance in natural infections remains pure speculation. Like the authors themselves admit, it is perfectly possible that some of these RNAs are just by-products of normal transcription and that some of the predicted proteins are not expressed or rapidly degraded [58, 61]. There is a pressing need to perform basic studies into the functional roles of the newly identified viral RNAs. From the collected data it is, however, clear that the combination of alternative splicing, overlapping transcription with alternative 5' ends, antisense transcription, large and small non-coding RNAs and miRNAs offer a complex regulatory framework through which HCMV can fine-tune its infectious cycle. It is also clear that mutational studies should be designed and interpreted with extensive care taking into account these added levels of complexity.

CONCLUSIONS AND OUTLOOK

Although more than 20 years have passed since the first full genome sequence of an HCMV strain has been published, we are still only beginning to comprehend the complete coding capacity of this ubiquitous virus. The observation that *in vitro* passaged viral strains undergo genetic adaptations led to the realization that focus should be put on characterizing recent clinical isolates. The new generation of sequencing technologies emerged over the past decade, offer the possibilities to analyse a wide variety of clinical strains to characterize the variability in coding capacity among different HCMV isolates, to determine the importance of viral mutants in different patient groups and to potentially associate genetic diversity with viral pathogenicity. While these new technologies offer great possibilities, several challenges remain. The purification and enrichment of viral DNA from recent clinical isolates is still a major obstacle to the high-throughput application of NGS technologies. Protocols based on DNase-mediated degradation of non-viral DNA [68] or target enrichment through probe hybridisation [69] could provide solutions. In addition, the evolutionary dynamics of intrahost viral populations should be studied in greater detail to assess potential implications of population bottlenecks and selective pressures. Meanwhile, NGS technologies are constantly evolving and new developments in single-molecule sequencing will open up a completely new set of possibilities like the elucidation of DNA modifications [70]. These technological developments will have to go hand-in-hand with advances in bioinformatics algorithms, which is just as major a challenge [71]. Finally, an enormous effort lies ahead in dissecting the transcriptional regulatory features that are being unveiled by genome-wide transcriptome analyses. These endeavours will certainly offer new insights in HCMV pathogenicity and could potentially lead to the development of new antiviral therapies.

ACKNOWLEDGEMENTS

SS and PM are supported by the Research Foundation Flanders (FWO – 'Fonds voor Wetenschappelijk Onderzoek, Vlaanderen').

REFERENCES

1. Davison, A. J. 2010, Vet. Microbiol., 143, 52.
2. Cannon, M. J., Schmid, D. S. and Hyde, T. B. 2010, Rev. Med. Virol., 20, 202.
3. Britt, W. 2008, Curr. Top. Microbiol. Immunol., 325, 417.
4. Kenneson, A. and Cannon, M. J. 2007, Rev. Med. Virol., 17, 253.
5. Murphy, E. and Shenk, T. 2008, Curr. Top. Microbiol. Immunol., 325, 1.
6. Chee, M. S., Bankier, A. T., Beck, S., Bohni, R., Brown, C. M., Cerny, R., Horsnell, T., Hutchison, C. A. 3rd, Kouzarides, T. and Martignetti, J. A. 1990, Curr. Top. Microbiol. Immunol., 154, 125.
7. Bankier, A. T., Beck, S., Bohni, R., Brown, C. M., Cerny, R., Chee, M. S., Hutchison, C. A. 3rd, Kouzarides, T., Martignetti, J. A. and Preddie, E. 1991, DNA Seq., 2, 1.
8. Cha, T. A., Tom, E., Kemble, G. W., Duke, G. M., Mocarski, E. S. and Spaete, R. R. 1996, J. Virol., 70, 78.
9. Prichard, M. N., Penfold, M. E., Duke, G. M., Spaete, R. R. and Kemble, G. W. 2001, Rev. Med. Virol., 11, 191.
10. Davison, A. J., Dolan, A., Akter, P., Addison, C., Dargan, D. J., Alcendor, D. J., McGeoch, D. J. and Hayward, G. S. 2003, J. Gen. Virol., 84, 17.

11. Murphy, E., Rigoutsos, I., Shibuya, T. and Shenk, T. E. 2003, Proc. Natl. Acad. Sci. USA, 100, 13585.
12. Murphy, E., Yu, D., Grimwood, J., Schmutz, J., Dickson, M., Jarvis, M. A., Hahn, G., Nelson, J. A., Myers, R. M. and Shenk, T. E. 2003, Proc. Natl. Acad. Sci. USA, 100, 14976.
13. Dolan, A., Cunningham, C., Hector, R. D., Hassan-Walker, A. F., Lee, L., Addison, C., Dargan, D. J., McGeoch, D. J., Gatherer, D., Emery, V. C., Griffiths, P. D., Sinzger, C., McSharry, B. P., Wilkinson, G. W. and Davison, A. J. 2004, J. Gen. Virol., 85, 1301.
14. Bradley, A. J., Lurain, N. S., Ghazal, P., Trivedi, U., Cunningham, C., Baluchova, K., Gatherer, D., Wilkinson, G. W., Dargan, D. J. and Davison, A. J. 2009, J. Gen. Virol., 90, 2375.
15. Revello, M. G. and Gerna, G. 2010, Rev. Med. Virol., 20, 136.
16. Stanton, R. J., Baluchova, K., Dargan, D. J., Cunningham, C., Sheehy, O., Seirafian, S., McSharry, B. P., Neale, M. L., Davies, J. A., Tomasec, P., Davison, A. J. and Wilkinson, G. W. 2010, J. Clin. Invest., 120, 3191.
17. Dargan, D. J., Douglas, E., Cunningham, C., Jamieson, F., Stanton, R. J., Baluchova, K., McSharry, B. P., Tomasec, P., Emery, V. C., Percivalle, E., Sarasini, A., Gerna, G., Wilkinson, G. W. and Davison, A. J. 2010, J. Gen. Virol., 91, 1535.
18. Mardis, E. R. 2008, Annu. Rev. Genomics. Hum. Genet., 9, 387.
19. Cunningham, C., Gatherer, D., Hilfrich, B., Baluchova, K., Dargan, D. J., Thomson, M., Griffiths, P. D., Wilkinson, G. W., Schulz, T. F. and Davison, A. J. 2010, J. Gen. Virol., 91, 605.
20. Jung, G. S., Kim, Y. Y., Kim, J. I., Ji, G. Y., Jeon, J. S., Yoon, H. W., Lee, G. C., Ahn, J. H., Lee, K. M. and Lee, C. H. 2011, Virus Res., 156, 113.
21. Görzer, I., Guelly, C., Trajanoski, S. and Puchhammer-Stockl, E. 2010, J. Virol., 84, 7195.
22. Renzette, N., Bhattacharjee, B., Jensen, J. D., Gibson, L. and Kowalik, T. F. 2011, PLoS Pathog., 7, e1001344.
23. Chou, S. W. and Dennison, K. M. 1991, J. Infect. Dis., 163, 1229.
24. Pignatelli, S., Dal Monte, P. and Landini, M. P. 2001, J. Gen. Virol., 82, 2777.
25. Rasmussen, L., Geissler, A., Cowan, C., Chase, A. and Winters, M. 2002, J. Virol., 76, 10841.
26. Lurain, N. S., Kapell, K. S., Huang, D. D., Short, J. A., Paintsil, J., Winkfield, E., Benedict, C. A., Ware, C. F. and Bremer, J. W. 1999, J. Virol., 73, 10040.
27. Lurain, N. S., Fox, A. M., Lichy, H. M., Bhorade, S. M., Ware, C. F., Huang, D. D., Kwan, S. P., Garrity, E. R. and Chou, S. 2006, Virology Journal, 3, 4.
28. Qi, Y., Mao, Z. Q., Ruan, Q., He, R., Ma, Y. P., Sun, Z. R., Ji, Y. H. and Huang, Y. 2006, J. Med. Virol., 78, 517.
29. Davison, A. J., Akter, P., Cunningham, C., Dolan, A., Addison, C., Dargan, D. J., Hassan-Walker, A. F., Emery, V. C., Griffiths, P. D. and Wilkinson, G. W. 2003, J. Gen. Virol., 84, 657.
30. Hitomi, S., Kozuka-Hata, H., Chen, Z., Sugano, S., Yamaguchi, N. and Watanabe, S. 1997, Arch. Virol., 142, 1407.
31. Bar, M., Shannon-Lowe, C. and Geballe, A. P. 2001, J. Infect. Dis., 183, 218.
32. Sekulin, K., Gorzer, I., Heiss-Czedik, D. and Puchhammer-Stockl, E. 2007, Virus Genes, 35, 577.
33. Stanton, R., Westmoreland, D., Fox, J. D., Davison, A. J. and Wilkinson, G. W. 2005, J. Med. Virol., 75, 42.
34. Pignatelli, S., Dal Monte, P., Rossini, G., Chou, S., Gojobori, T., Hanada, K., Guo, J. J., Rawlinson, W., Britt, W., Mach, M. and Landini, M. P. 2003, J. Gen. Virol., 84, 647.
35. Bradley, A. J., Kovacs, I. J., Gatherer, D., Dargan, D. J., Alkharsah, K. R., Chan, P. K., Carman, W. F., Dedicoat, M., Emery, V. C., Geddes, C. C., Gerna, G., Ben-Ismaeil, B., Kaye, S., McGregor, A., Moss, P. A., Pusztai, R., Rawlinson, W. D., Scott, G. M., Wilkinson, G. W., Schulz, T. F. and Davison, A. J. 2008, J. Med. Virol., 80, 1615.
36. Bates, M., Monze, M., Bima, H., Kapambwe, M., Kasolo, F. C. and Gompels, U. A. 2008, Virology, 382, 28.
37. Puchhammer-Stockl, E. and Gorzer, I. 2011, Future Virology, 6, 259.

38. McGeoch, D. J., Rixon, F. J. and Davison, A. J. 2006, Virus Res., 117, 90.
39. Rasmussen, L., Geissler, A. and Winters, M. 2003, J. Infect. Dis., 187, 809.
40. Chou, S. W. 1989, J. Infect. Dis., 160, 11.
41. Faure-Della Corte, M., Samot, J., Garrigue, I., Magnin, N., Reigadas, S., Couzi, L., Dromer, C., Velly, J. F., Dechanet-Merville, J., Fleury, H. J. and Lafon, M. E. 2010, J. Clin. Virol., 47, 161.
42. Torok-Storb, B., Boeckh, M., Hoy, C., Leisenring, W., Myerson, D. and Gooley, T. 1997, Blood, 90, 2097.
43. Rossini, G., Pignatelli, S., Dal Monte, P., Camozzi, D., Lazzarotto, T., Gabrielli, L., Gatto, M. R. and Landini, M. P. 2005, Microbes Infect, 7, 890.
44. Pignatelli, S., Lazzarotto, T., Gatto, M. R., Dal Monte, P., Landini, M. P., Faldella, G. and Lanari, M. 2010, Clin. Infect. Dis., 51, 33.
45. Heo, J., Petheram, S., Demmler, G., Murph, J. R., Adler, S. P., Bale, J. and Sparer, T. E. 2008, Virology, 378, 86.
46. Pignatelli, S., Dal Monte, P., Rossini, G. and Landini, M. P. 2004, Rev. Med. Virol., 14, 383.
47. Puchhammer-Stockl, E. and Gorzer, I. 2006, J. Clin. Virol., 36, 239.
48. Gorzer, I., Kerschner, H., Redlberger-Fritz, M. and Puchhammer-Stockl, E. 2010, J. Clin. Virol., 48, 100.
49. Ross, S. A., Novak, Z., Pati, S., Patro, R. K., Blumenthal, J., Danthuluri, V. R., Ahmed, A., Michaels, M. G., Sanchez, P. J., Bernstein, D. I., Tolan, R. W., Palmer, A. L., Britt, W. J., Fowler, K. B. and Boppana, S. B. 2011, J. Infect. Dis., 204, 1003.
50. Cicin-Sain, L., Podlech, J., Messerle, M., Reddehase, M. J. and Koszinowski, U. H. 2005, J. Virol., 79, 9492.
51. Humar, A., Kumar, D., Gilbert, C. and Boivin, G. 2003, J. Infect. Dis., 188, 581.
52. Coaquette, A., Bourgeois, A., Dirand, C., Varin, A., Chen, W. and Herbein, G. 2004, Clin. Infect. Dis., 39, 155.
53. Puchhammer-Stockl, E., Gorzer, I., Zoufaly, A., Jaksch, P., Bauer, C. C., Klepetko, W. and Popow-Kraupp, T. 2006, Transplantation, 81, 187.
54. Manuel, O., Asberg, A., Pang, X., Rollag, H., Emery, V. C., Preiksaitis, J. K., Kumar, D., Pescovitz, M. D., Bignamini, A. A., Hartmann, A., Jardine, A. G. and Humar, A. 2009, Clin. Infect. Dis., 49, 1160.
55. Bhattacharjee, B., Renzette, N. and Kowalik, T. F. 2012, J. Virol., 86, 6815.
56. Ma, Y., Wang, N., Li, M., Gao, S., Wang, L., Zheng, B., Qi, Y. and Ruan, Q. 2012, Future Microbiol., 7, 577.
57. Zhang, G., Raghavan, B., Kotur, M., Cheatham, J., Sedmak, D., Cook, C., Waldman, J. and Trgovcich, J. 2007, J. Virol., 81, 11267.
58. Gatherer, D., Seirafian, S., Cunningham, C., Holton, M., Dargan, D. J., Baluchova, K., Hector, R. D., Galbraith, J., Herzyk, P., Wilkinson, G. W. and Davison, A. J. 2011, Proc. Natl. Acad. Sci. USA, 108, 19755.
59. Su, W. Y., Xiong, H. and Fang, J. Y. 2010, Biochem. Biophys. Res. Commun., 396, 177.
60. Reeves, M. B., Davies, A. A., McSharry, B. P., Wilkinson, G. W. and Sinclair, J. H. 2007, Science, 316, 1345.
61. Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V. T., Hein, M. Y., Huang, S. X., Ma, M., Shen, B., Qian, S. B., Hengel, H., Mann, M., Ingolia, N. T. and Weissman, J. S. 2012, Science, 338, 1088.
62. Pfeffer, S., Zavolan, M., Grasser, F. A., Chien, M., Russo, J. J., Ju, J., John, B., Enright, A. J., Marks, D., Sander, C. and Tuschl, T. 2004, Science, 304, 734.
63. Bartel, D. P. 2009, Cell, 136, 215.
64. Dunn, W., Trang, P., Zhong, Q., Yang, E., van Belle, C. and Liu, F. 2005, Cell. Microbiol., 7, 1684.
65. Grey, F., Antoniewicz, A., Allen, E., Saugstad, J., McShea, A., Carrington, J. C. and Nelson, J. 2005, J. Virol., 79, 12095.
66. Tuddenham, L. and Pfeffer, S. 2011, Biochim. Biophys. Acta, 1809, 613.
67. Stark, T. J., Arnold, J. D., Spector, D. H. and Yeo, G. W. 2012, J. Virol., 86, 226.
68. Volkening, J. D. and Spatz, S. J. 2009, J. Virol. Methods, 157, 55.
69. Depledge, D. P., Palser, A. L., Watson, S. J., Lai, I. Y., Gray, E. R., Grant, P., Kanda, R. K., Leproust, E., Kellam, P. and Breuer, J. 2011, PLoS One, 6, e27805.

70. Korlach, J. and Turner, S. W. 2012, Curr. Opin. Struct. Biol., 22, 251.
71. McPherson, J. D. 2009, Nat. Methods, 6, S2.
72. Nei, M. and Gojobori, T. 1986, Mol. Biol. Evol., 3, 418.
73. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. 2011, Mol. Biol. Evol., 28, 2731.
74. Dargan, D. J., Jamieson, F. E., MacLean, J., Dolan, A., Addison, C. and McGeoch, D. J. 1997, J. Virol., 71, 9833.
75. Dunn, W., Chou, C., Li, H., Hai, R., Patterson, D., Stolc, V., Zhu, H. and Liu, F. 2003, Proc. Natl. Acad. Sci. USA, 100, 14223.
76. Sinzger, C., Hahn, G., Digel, M., Katona, R., Sampaio, K. L., Messerle, M., Hengel, H., Koszinowski, U., Brune, W. and Adler, B. 2008, J. Gen. Virol., 89, 359.